UNIVERSITY OF CAMBRIDGE

OSC
Office of Scholarly Communication

Text and Data Mining: One Year On

Dr Danny Kingsley and Dr Debbie Hansen
Office of Scholarly Communication
dak45@cam.ac.uk, dh554@cam.ac.uk

Outline of this meeting

Text and Data Mining

The past year

Currently in progress

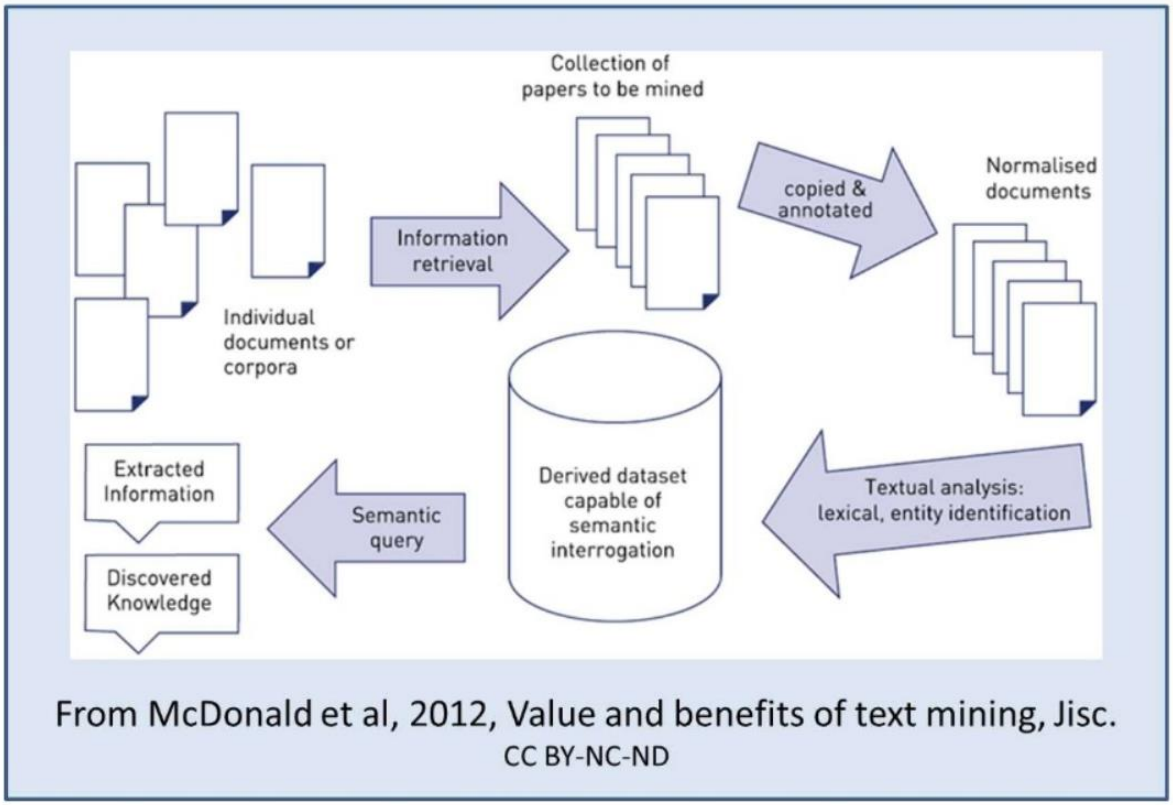
Discussion

Towards developing a TDM library service

OSC
Office of Scholarly Communication

What is Text and Data Mining (TDM)?

TDM is the process of using digital techniques to explore large collections of machine-readable material, extracting new datasets and, by analysis, finding out new information about a topic



From McDonald et al, 2012, Value and benefits of text mining, Jisc.

OSC
Office of Scholarly Communication

The past year

Text and Data Mining Services: What can Cambridge Libraries Offer? February 2017

~30 attendees, mainly library staff
Wide-ranging discussion and the following was agreed:

What could a Future Library TDM Support Service include?

- ACCESS to data from our own collections
- ADVICE on
  - Legal issues
  - What publishers allow
  - What data sets are available
  - What tools are available
- REGISTERS on
  - Data provided for mining
  - TDM projects
- FOSTER agreements with publishers

Next Steps

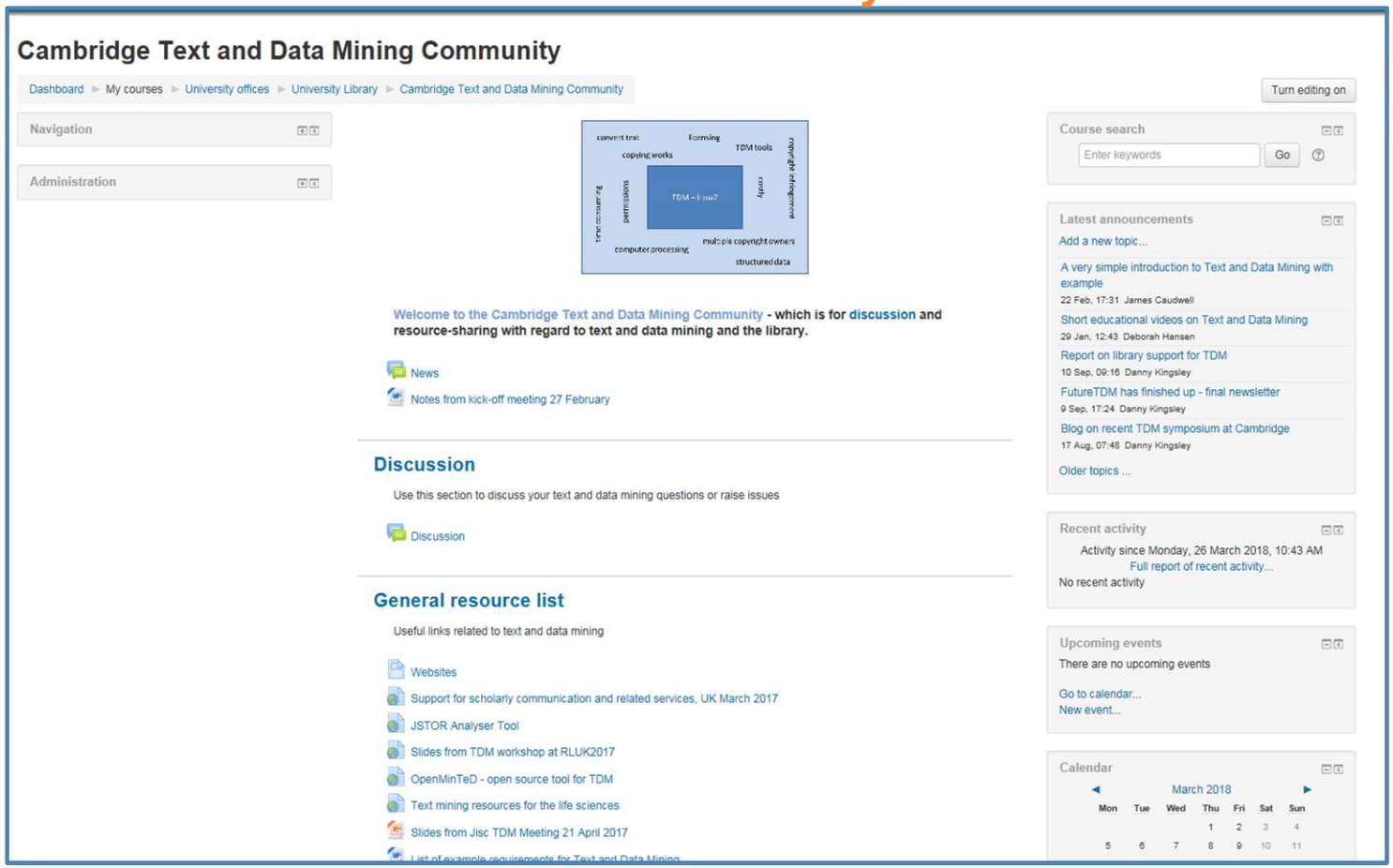
- MORE INFORMATION e.g.
  - Demonstrations – what TDM looks like
  - Case studies – that we can publish on-line
- WOULD BE USEFUL:
  - A Moodle group: TDM discussions/resources in one place
  - A TDM area on the library website
- WORKING GROUP: to help develop a University of Cambridge TDM service

(Kingsley, 2017, Notes from kick-off meeting 27 February)

OSC
Office of Scholarly Communication

The past year

TDM Moodle for communication within the community



https://www.vle.cam.ac.uk/course/view.php?id=134922

OSC
Office of Scholarly Communication

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Workshop at RLUK, 2017
Dr Danny Kingsley, Dr Debbie Hansen – University of Cambridge
Anna Vernon – Jisc
British Library 9th March 2017

- 19 mainly librarians from large/small/other Universities/research institutions; couple publishers.
- Discussion 1: Talk about any experiences you have had with TDM
- Discussion 2: Why your organisation is not actively supporting TDM or if it is, what are the top challenges you face?
- Discussion 3: If we were to draft a statement for a Service Level Agreement for publishers to assure us that if the activity is legal we will be reinstated within 1 hour (or something like that), what are the issues if we did this?

OSC
Office of Scholarly Communication

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Discussion 1: Talk about any experiences you have had with TDM

Following this discussion it was agreed that there was a need for:

- UNDERSTANDING regarding licensing
- A MECHANISM for advice - where to go within an institution and a publisher
- POLICY - development of procedures for handling TDM related requests
- ADDRESSING researcher behaviour – academics are not always concerned by copyright

OSC
Office of Scholarly Communication

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Discussion 2: Why your organisation is not actively supporting TDM? If it is, what are the top challenges you face?

Why not supporting?

- Practical reasons
  - challenges of handling physical media
  - risk of lockout
- Lack of demand
  - Not getting enquiries
  - Not much call
- Who is responsible?
  - Policy needed – issues not raised at academic level
  - Library service – how to scale up from individual queries?
  - Not joined up as organisations

Challenges:

- Understanding at the content-owner level
  - Library ensure people know their responsibilities
    - E.g. not for commercial use
  - Even if intrusive on the research process
- Time
  - Example of TDM contract between publisher and researchers
    - Took 2 years to finalise

OSC
Office of Scholarly Communication

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Discussion 3: If we were to draft a statement for a Service Level Agreement for publishers to assure us that if the activity is legal we will be reinstated within 1 hour (or something like that), what are the issues if we did this?

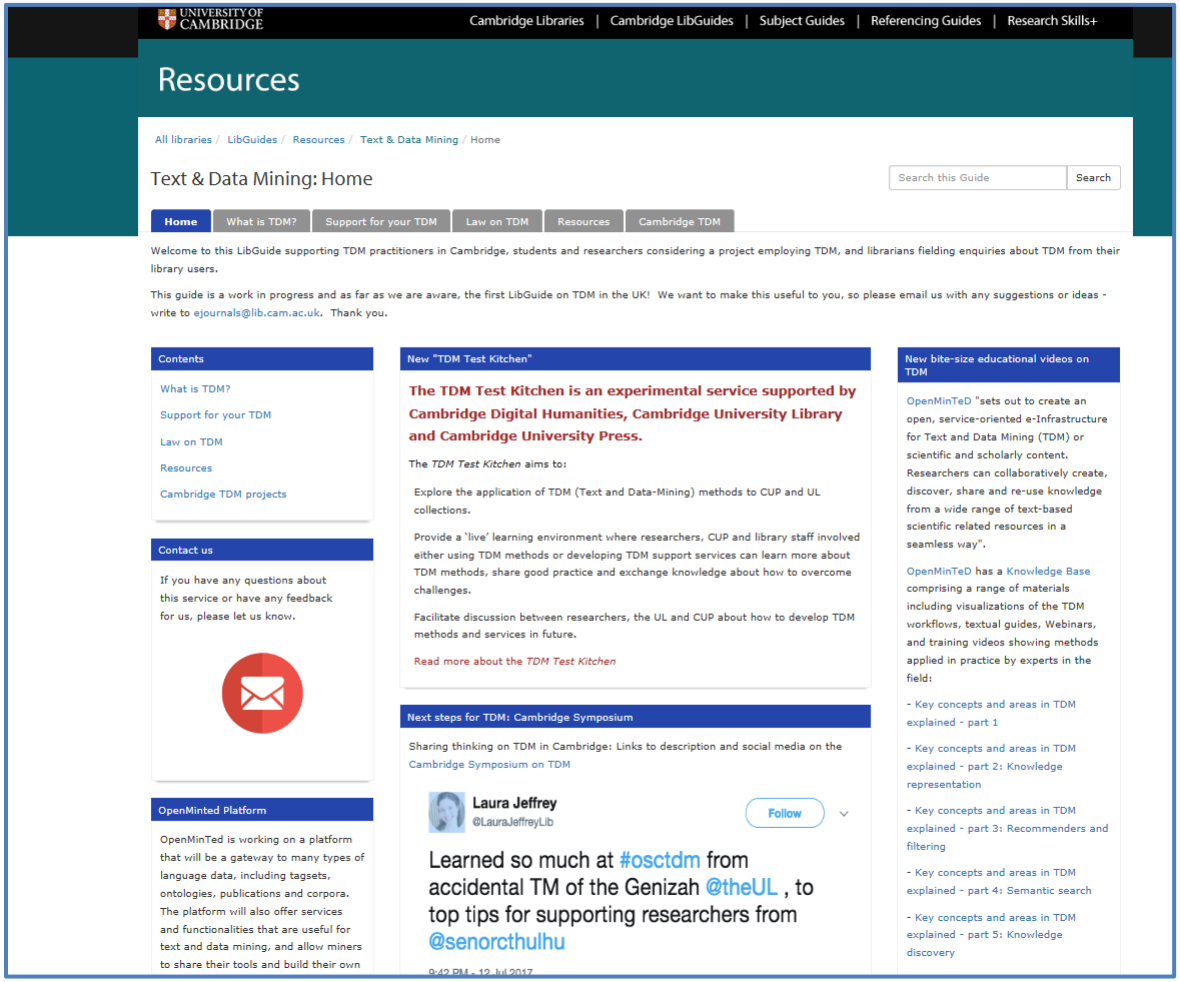
AGREED EXPECTATIONS

- Don't cut us off! Have a conversation first (and if you want to cut us off - prove there are all these activities happening in the UK)
- If you do cut us off and it turns out to be legitimate then we expect compensation for the time we were cut off
- Mechanisms for TDM where certain behaviours are expected - built into separate licensing agreements for TDM

OSC
Office of Scholarly Communication

The past year

Text and Data Mining LibGuide



Developed by eResources team

http://libguides.cam.ac.uk/tdm

OSC
Office of Scholarly Communication

The past year

University of Cambridge TDM Symposium

12th July 2017

OBJECTIVE: To provide as much information as possible to the attendees regarding TDM

WHO CAME?

- postgraduate students
- early career researchers
- senior researchers
- administrative staff
- librarians
- publishers

WHAT WAS THERE TO LEARN?

- Talks and Show and Tell sessions, e.g.
  - Keynote from Kiera McNeice (Future TDM)
    - TDM overview, what the barriers are
  - TDM in action
    - E.g. TDM and the Cairo Genizah
  - Tools e.g. ChemDataExtractor, ContentMine
  - How a publisher supports TDM (PLoS)
  - How librarians support TDM
- Discussion: The future of TDM in policy

https://unlockingresearch-blog.lib.cam.ac.uk/?p=1505

OSC
Office of Scholarly Communication

The past year

University of Cambridge TDM Symposium

Discussion outcomes

Why not more TDM happening in UK?

- Legal challenges
- Recruiting and training staff
- Lack of institutional TDM policies
- Lack of institutional/governmental leadership
- Publisher issues
  - What happens when a publisher cuts off access
  - Convoluted process to reinstate access
  - Impractical to inform publishers when TDM to take place
  - Threats from publishers if download a lot of material
  - Will TDM drive price increases?
- Lack of understanding of how the different TDM stakeholders work

NEXT STEPS

- Hands-on sessions
  - Discipline specific?
- Stakeholder communication
  - Publishers
  - Librarians
  - Researchers
  - Legal expertise

OSC
Office of Scholarly Communication

The past year

Work with FutureTDM/OpenMinTed/LERU

Involvement with EU TDM developments

INCLUDES:

- Review of TDM Guidelines: outputs from the Future TDM Project
  - https://www.futuretdm.eu/practitioner-guidelines/
- University of Cambridge Librarian support for EU TDM Copyright Exception
  - Letters to our MEPs on the European Parliament Committee on Legal Affairs and our national Government minister for science calling for TDM Exception in legislative proposals of European Parliament and Council of the European Union to be strengthened and widened
- Future TDM Symposium (at iDSC '17, June 2017, Salzburg, Austria)
  - Panel discussion on Universities, TDM and the need for strategic thinking on educating researchers

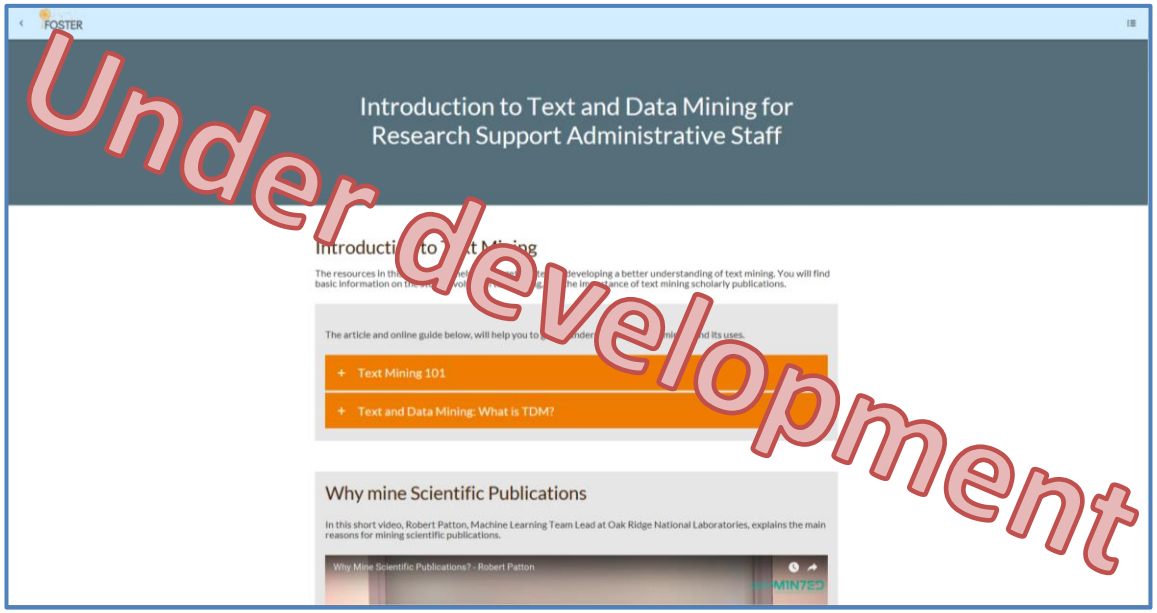
OSC
Office of Scholarly Communication

Current activities in progress

OSC Working with OpenMinTeD: Text and Data Mining On-line Course

OpenMinTeD Foster on-line course

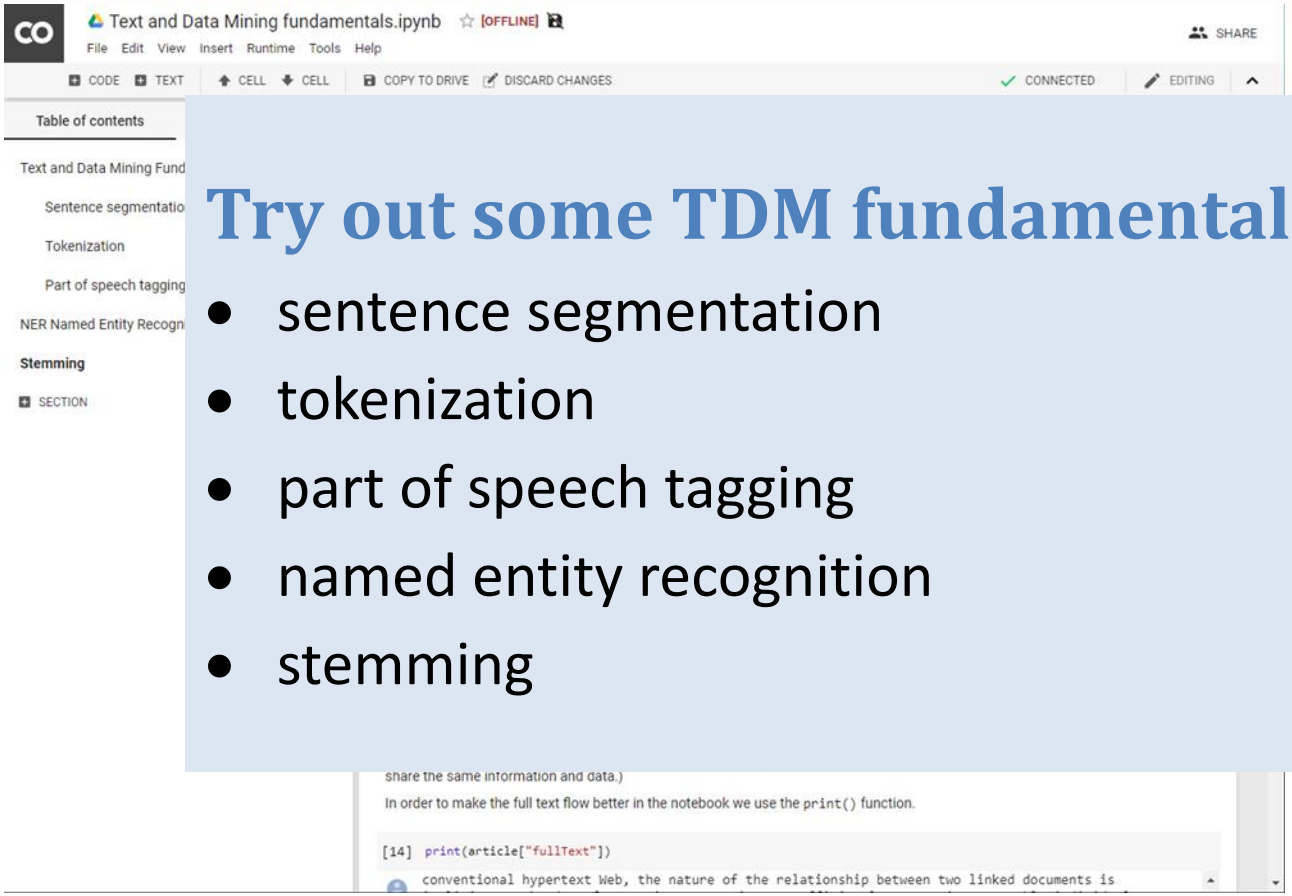
- Reading material
- Videos
- Quizzes
- Practical activities
- Glossaries



OSC
Office of Scholarly Communication

Current activities in progress

TDM on-line course: practical activities



Try out some TDM fundamentals such as

- sentence segmentation
- tokenization
- part of speech tagging
- named entity recognition
- stemming

OSC
Office of Scholarly Communication

Current activities in progress

TDM on-line course: practical activities

More advanced practical activity

- Mine selection of CORE articles
- Create a Word Cloud

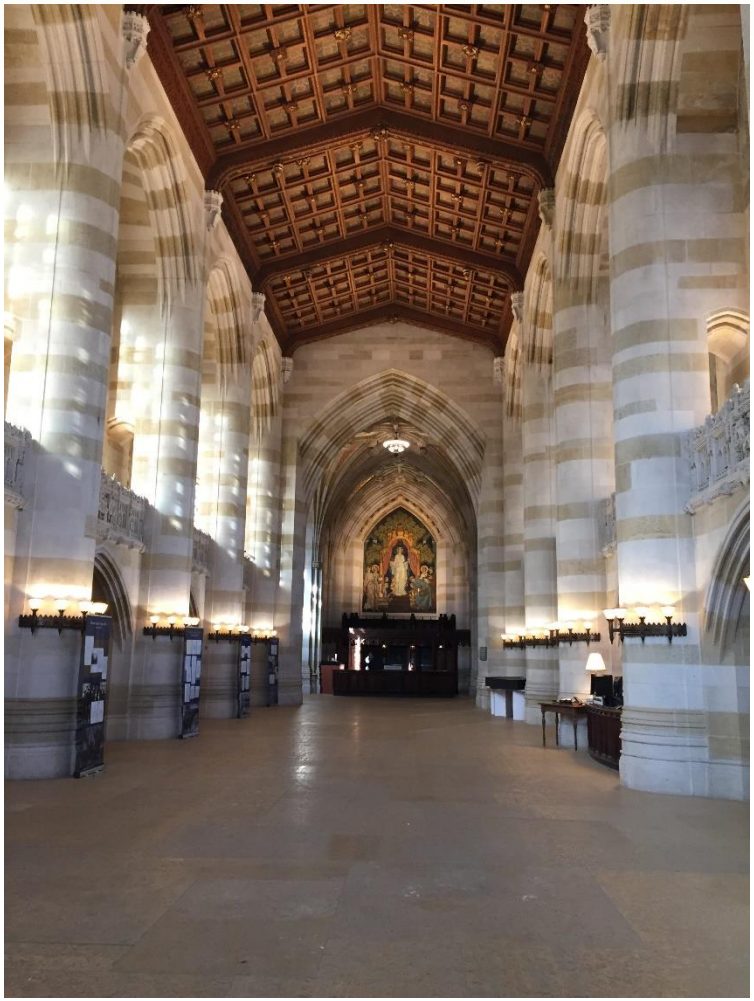
Office of Scholarly Communication

The past year

Dr Danny Kingsley visit to Yale: January 2018

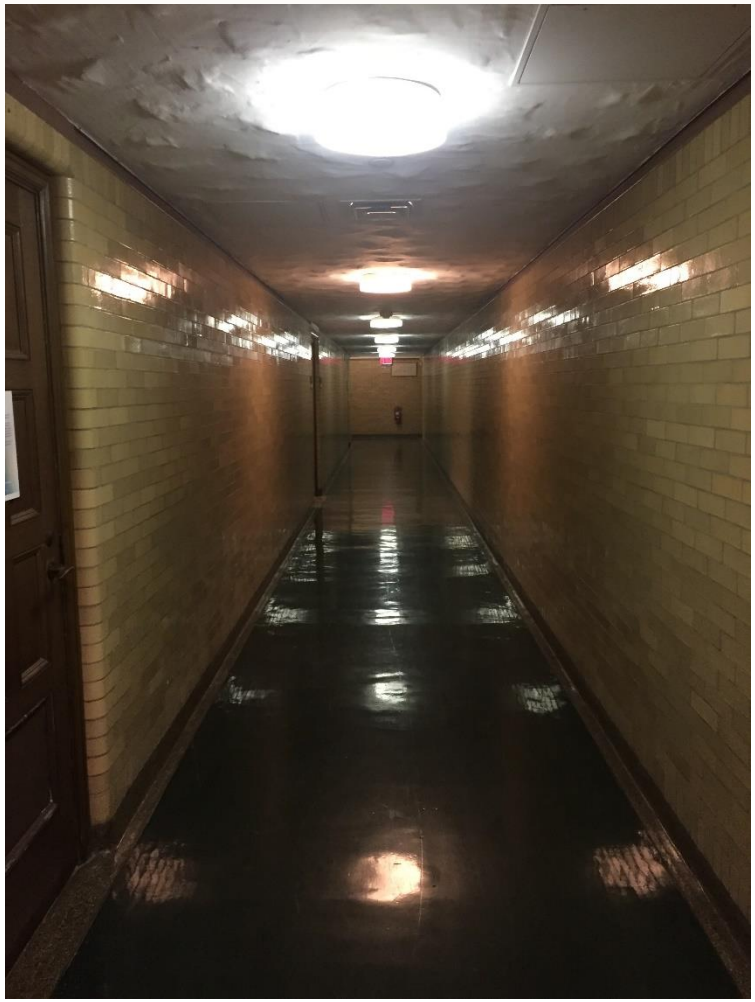
OSC
Office of Scholarly Communication

The hallowed halls of learning



Automated barriers to access stacks

Corridors in the Library



If you have a Yale card you can go through the barrier and are free to walk around here.

The DH Lab

Digital Humanities Lab

One Director, a UX person, an outreach person and two postdocs. Currently coming to end of four year grant, library is looking for $ to get them on staff.

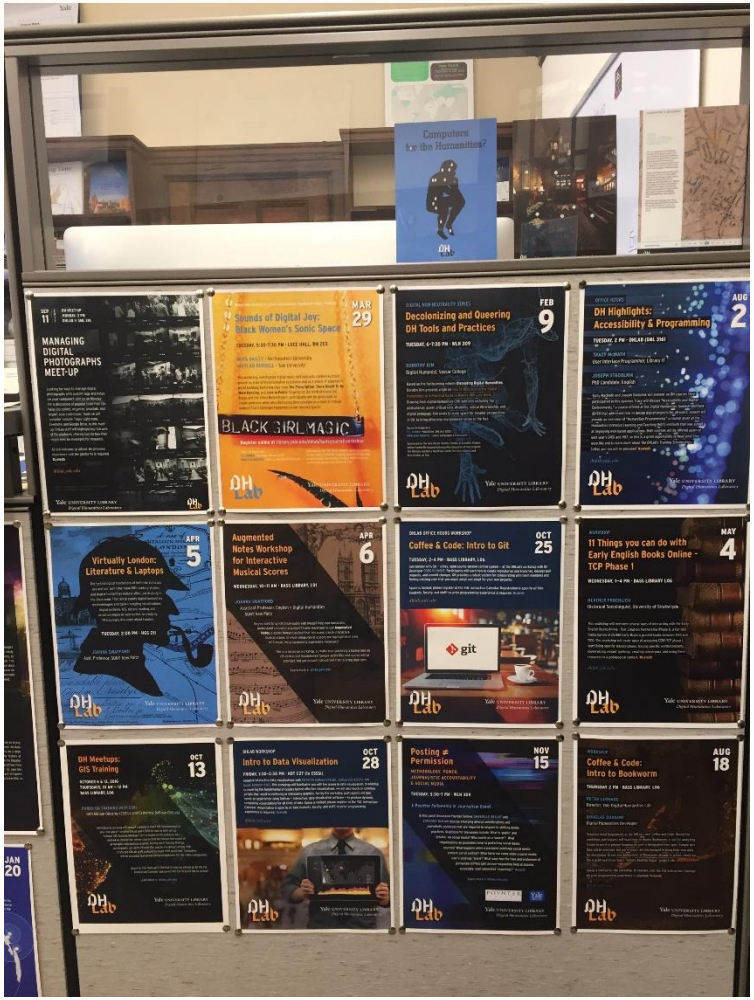
Moving to a bigger dedicated space on 1st floor in June with secure central glass box for sensitive data work

Digital Humanities Lab

These seats are for any student who wants to come in and work.
The red boxes contain data sets on hard drives that can be mined.

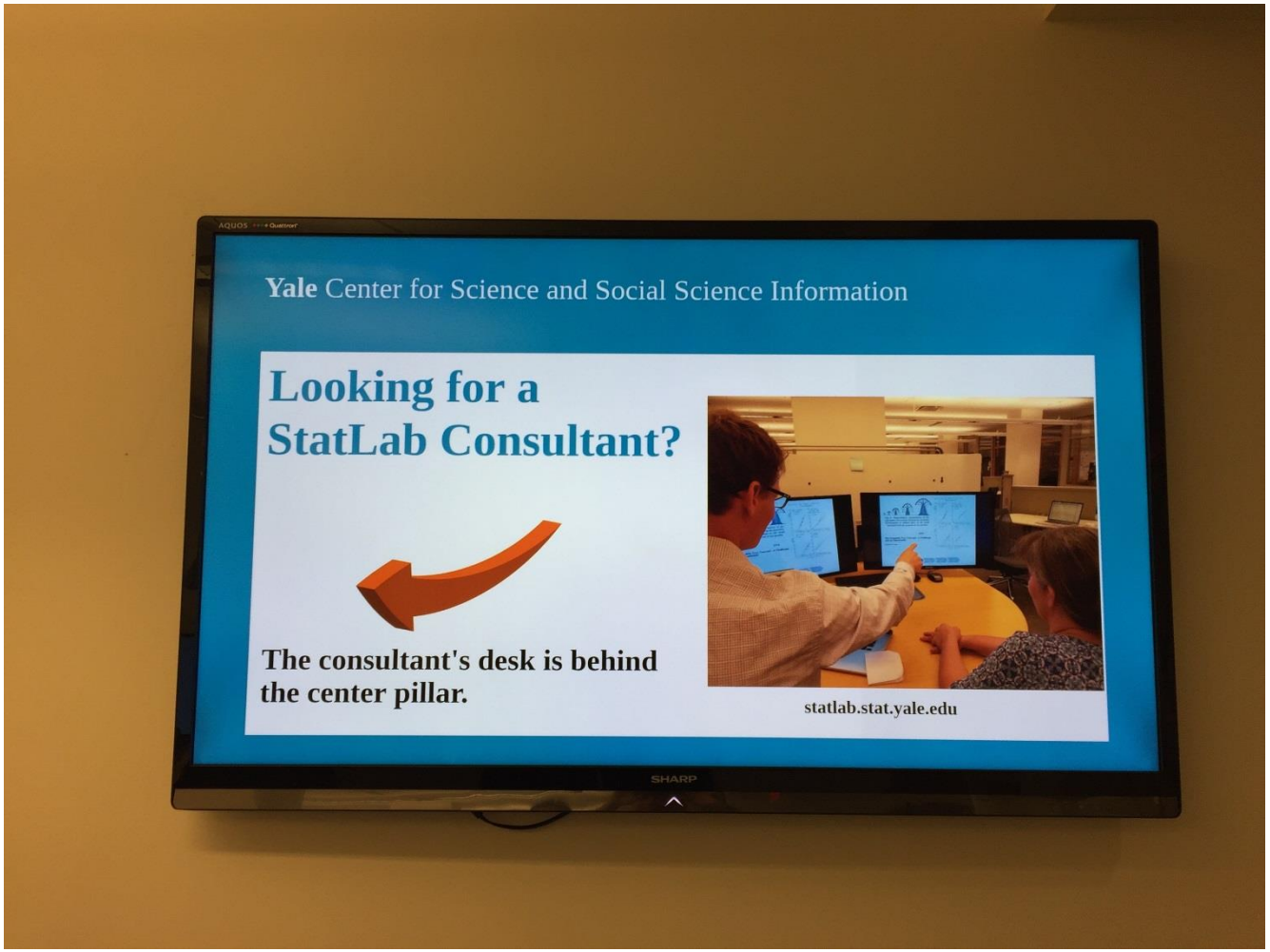
OSC
Office of Scholarly Communication

Digital Humanities Lab



An example of the events they run regularly

StatLab is part of the DH offering



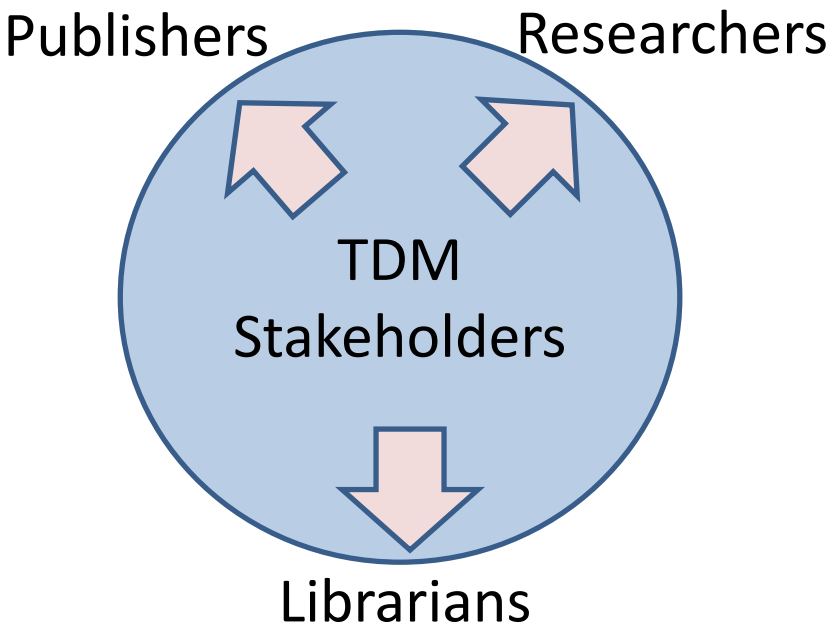
This IS the StatLab

A couple of postdocs are employed to man the desk

Current activities in progress

The TDM Test Kitchen

Pilot Project



OSC
Office of Scholarly Communication

Current activities in progress

The TDM Test Kitchen

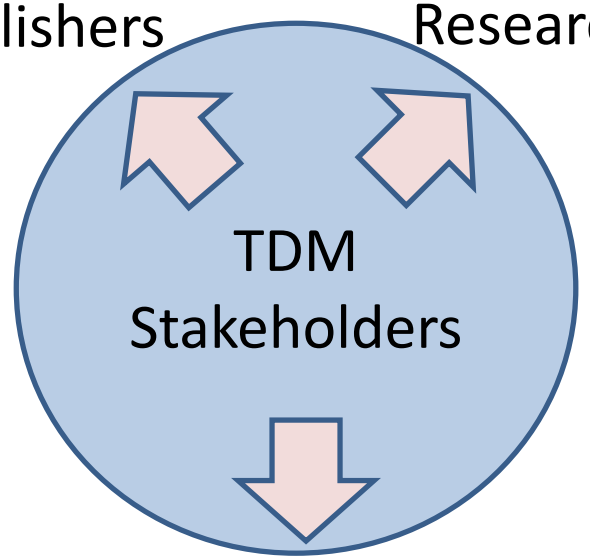
Pilot Project

Cambridge University Press

Publishers

Cambridge Digital Humanities

Researchers



Librarians

University Library

OSC
Office of Scholarly Communication

Current activities in progress

The TDM Test Kitchen

Pilot Project

https://www.digitalhumanities.cam.ac.uk/Methods/tdm-testkitchen

Experimental service aims to:

- Explore application of TDM methods to CUP and UL collections
- Provide a 'live' learning environment
  - Methods, developing support services, share good practice, overcoming challenges
- Facilitate discussion about how to develop TDM methods and services in the future

OSC
Office of Scholarly Communication

Current activities in progress

The TDM Test Kitchen

Pilot Project

https://www.digitalhumanities.cam.ac.uk/Methods/tdm-testkitchen

Outline of Pilot Phase:

- Run from Jan – Jul 2018
- Open to Cambridge University graduate students and staff
- Advice provided to participants (from IP rights to data visualisation)
- Outputs – short report for website and presentation in Jun/Jul
- Create a set of case-studies which will inform future TDM service developments

OSC
Office of Scholarly Communication

Early lessons learned (from an observer)

Early queries and outcomes (UL):

- Investigation into a particular journal: some under subscription, some purchased, could get through a particular library?, obtained record of all payments to journal to date to check access rights, what about the pre-1950 content?.. (i.e. 1 journal, not straightforward)
- Is there a need to think about original digitization (OCR)?
- Where is the agreed corpus that is to be mined intended to be held?
- Do we need a dedicated Helpdesk? (TDM enquiry form on the TDM LibGuide at moment, access to this is eresources team)
- Requests made to publishers (**not just CUP**) for material for this pilot and agreement to supply copies on drives obtained :-
- It is emerging that there is often a cost associated with obtaining material for mining (O(£1000) per item) – from which budget can these costs be met?

OSC
Office of Scholarly Communication

Early lessons learned (from an observer)

Early queries and outcomes (publisher):

- Is the metadata and content in a form as required for the tools?
- Where is the agreed corpus that is to be mined intended to be held? A TDM-specific environment needed.
- Outputs for non-commercial research purposes and limited use of outputs from source material.
- Addendum to current licences.
- Content requested not always readily available on modern platforms which can hold up delivery time.

Early lessons learned (from an observer)

Early process (UL):

- Obtain and clarify request from researcher
- Investigate what available and rights and file format required
- Place request with publisher; publisher develops licence and terms and provides quote for costs
- Establish budget from which costs to be met; process payment and licence agreement if any
- Obtain material in agreed format
- Loan to researcher, who makes copy and returns original to UL
- Ensure researcher clear on terms of use.

OSC
Office of Scholarly Communication

Time for Discussion

Towards developing a TDM library service

Actions?

- Policy
- Procedures
- Costs
- Licensing
- Next pilot?

OSC
Office of Scholarly Communication